FME Enabled ETL Processes for Spatial and Attribute Data Analysis

Farhad Alam
Research Scholar,
Faculty of Computing and Information Technology,
Himalayan University
Arunachal Pradesh

Sanjay Pachauri, PhD
Head of Department,
CSE/IT,
IIMT College of Engineering
Greater Noida, U.P

ABSTRACT
ETL is a type of data integration that refers to the three steps (extract, transform, and load) used to blend data from multiple sources. It's often used to build a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse or other system. FME has a rich data model designed implement ETL. FME provides tremendous transformation functionality, resulting in output that can be much greater than the sum of the inputs, and allowing data to be transformed from one type to another. The current paper uses FME workbench and implement the concept of ETL using a case study where a private firm wants to integrate attribute and spatial information regarding its employee, filter the unnecessary information and finally implement business query regarding Monthly Travelling Allowance. The results establish ETL and FEM as interdisciplinary technological domain and backbone of the data warehouse architecture.

Keywords
Extract, Transform, and Load (ETL), Feature Manipulation Engine (FME), Keyhole Markup Language (KML), Attribute.

1. INTRODUCTION
The rapid proliferation of the Information and Communication in past couple of decades gave a new meaning to the phrase "information". Private Companies need to consider how to adopt and utilize real-time data and information into the fabric of their decision-making or risk falling behind their competitors. The challenge of extracting value from big data is similar in many ways to the age-old problem of distilling business intelligence from transactional data [1][2]. At the heart of this challenge is the process used to extract data from multiple sources, transform it to fit your analytical needs, and load it into a data warehouse for subsequent analysis, a process known as "Extract, Transform & Load" (ETL). A traditional ETL process extracts data from multiple sources, then cleanses, formats, and loads it into a data warehouse for analysis (Table 1) [3]. When the source data sets are large, fast, and unstructured, traditional ETL can become the bottleneck, because it is too complex to develop, too expensive to operate, and takes too long to execute. ETL process evolved and gradually took control over the Data Warehousing market to fulfill this requirement. Initially, organizations developed their own custom codes to perform the ETL activity which was referred as Hand-coded ETL process [4][5].

Table. 1 Various generations of ETL activity (Eckerson and White., 2003)

ERA	TITLE	SIGNIFICANCE
Early 1990	Hand Coded	Custom Codes (Hand Written)
1993-1997	1st Generation Tools	Code Based tools
1999-2001	2nd Generation Tools	Engine Based Tools
2003-2006	3rd Generation Tools	Efficient Tools
2007-2011	Parallel ETL Processing Tools	Intelligence Search and Optimization
2011-till Date	In Memory Computing	High speed processing and handling of huge datasets.

Current Study aims at developing concepts for the integration of spatial and attributes data sets. We have developed heterogeneous data extraction and implementation strategy in a FME workbench. FME Workbench provides a visual, flow chart-like environment for feature manipulation, consisting of a linked set of "transformers," each of which performs a particular data manipulation task.

2. EXTRACT-TRANSFORM-LOAD (ETL)
Enterprises have homogeneous and heterogeneous data sources, cannot rely over on line transaction processing (OLTP) or data warehouses for their services. Data Warehouses maintain aggregated format of data while OLTP incorporate all metadata corresponds to every instance of data [6]. OLTP maintains every single and short updates in transactions belongs to each data source, whereas data warehouse need long queries related to the large part of the database [7][8]. General framework to define the different instances of ETL processes are shown in the figure 1. This framework also depicts the cycle of ETL processing.

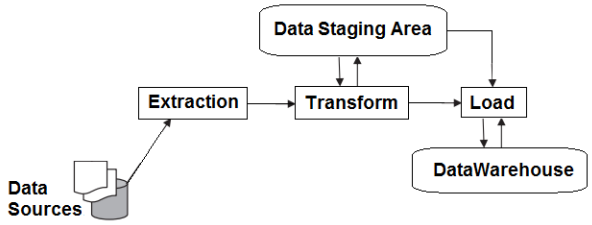

Fig. 1 General framework for ETL processes

ETL comes from Data Warehousing and stands for Extract-Transform-Load. Extract, Transform and Load refers to the process of extracting data from outside sources, transforms it to fit operational needs, loads it into the end target database, more specifically, operational data store or data warehouse [9][10].

- **Extraction:** The extraction part includes taking out the data from a variety of disparate source systems correctly is often the most challenging aspect of ETL.
- **Transformation:** The transformation step tends to make some cleaning and conforming on the incoming data to gain accurate data which is correct, complete, consistent, and unambiguous.
- **Load:** Load step involves loading the transformed data into one central repository where data is looked up for reporting purposes.

The meaningful framework for DW-based system refers to the capabilities of ETL processes to handling the problems that arise on maintenance phase such as data sources changes and data view refresh in DW.

3. FEATURE MANIPULATION ENGINE (FME)

Feature Manipulation Engine (FME) by Safe Software is operable over 300 formats and 400 transformers to allow interoperability between different data formats including excel, point cloud, 3D, raster, database, vector, KML and XML formats [11]

FME workbench 2016.1 is the space for concatenating reader and writer with series of transformers between for performing data conversion, transformation, integration and validation. Association of semantic information to the geometries of a model can also be carried out using (AttributeCreator) transformers.

FME Data inspector supports the visualization of data in wide range of formats. It also enables the verification of the color and linestyle features, the number of features and the layer information. Simple Thematic queries can also be performed using this Data inspector. It also supports database solutions including PostGIS, Oracle and Microsoft SQL Server.

4. MICROSOFT EXCEL

Microsoft Excel is a commercial spreadsheet application, written and distributed by Microsoft for Microsoft Windows and Mac OS X [12]. Excel allows us to enter quantitative data into an electronic spreadsheet to apply one or many mathematical computations. These computations ultimately convert that quantitative data into information. The information produced in Excel can be used to make decisions in both professional and personal contexts.

Small businesses often use Excel to create basic employee and resource schedules that can be color-coded and designed to automatically update as the schedules change. Excel is keep on growing as a popular choice for storing employee information that grows in detail over time, because we can add fields as they're needed without causing any problems with the existing data.

5. GOOGLE EARTH AND KML

Google Earth [13] is a programme that constructs pictures of the surface of our planet by downloading satellite data from a remote server. Since its release in June 2005, Google Earth has been bringing satellite images of our planet into our homes. Dropping a pin into a map of Google Earth on the mobile phone or tablet allows us to save a map position that we can return to at a later date, preventing the need of continuously searching for an often-referenced location. Once we find the location we want to save, dropping a pin or adding a place marker is a simple procedure.

KML is a file format used to display geographic data in an Earth browser, such as Google Earth, Google Maps, and Google Maps for Mobile [14]. KML uses a tag-based structure with nested elements and attributes and is based on the XML standard. One can create KML files with the Google Earth user interface, or you can use an XML or simple text editor to enter "raw" KML from scratch.

6. ESRI SHAPEFILES

A shapefile (.shp) is a simple geospatial format regulated by the Environmental Systems Research Institute (ESRI), used mainly for digitally storing a location or the features of an area. Shapefiles are the Department's preferred spatial format for permit area descriptions. ESRI shapefiles consist of three files. The first file (*.shp) contains the geography of each shape. The second file (*.shx) is an index file which contains record offsets. The third file (*.dbf) contains feature attributes with one record per feature.

7. CASE STUDY

7.1 Problem Statement

A company 'ComTech' wants to provide monthly traveling allowance to its employees on the basis of their shortest distance they travel to reach company office and their basic salary. This problem has a heterogeneous nature and depends up the spatial and attribute information related to each employee. Company holds employee record in the form of attribute information stored in the database while employee location as kml file holds latitude and longitude specific to each employee home location. Now the problem is to incorporate these two information (Attribute and Spatial) join them using unique employee id, filter them to select attribute important to perform business logic and ultimately query to decide travelling allowance for each employee.

7.2 Proposed Framework

A detailed framework is presented in the figure 2. Feature Manipulation Engine (FME) is placed at the center of the system and empRecord dataset prepared using Excel and empLocation dataset prepaed using Google Earth (Kml) is supplied as the input. FME perform data extraction and transformation with predefined filtering rule and the refined attribute with their location (latitude and longitude) is placed in the form of vector file (shp).

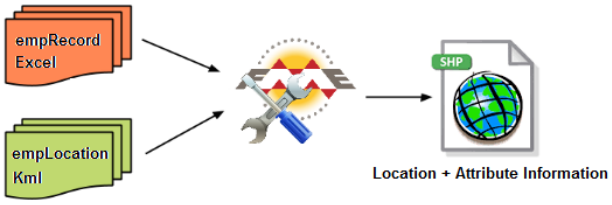

Fig. 2 Proposed framework

Studies uses this framework to collect existing employee information with a set of attribute written in excel and integrate it with the employee home information in terms of latitude and longitude written in Kml using Google earth. These two dataset combined and filtered to extract the useful information. The desired information holding employee Id, Name and Contact details with the home locations are placed in vector dataset (shp file).

7.3 Framework Implementation

As discussed in the problem statement a company named ComTech wants to use their employee's home address with their previously specified information to achieve various business outcomes. Employee information that is available with the company is specified in the table 5. Employee id, First Name, Last Name, Designation, Date of Birth (DOB), Mobile No and Email are the attribute under which this information is specified. Now the company collected the home latitude and longitude for each employee using Google earth and finally exported as the Kml file. One instance of that Kml file is presented in the code 1. Employee attribute information in terms of excel records and address in terms of coordinate now required to be clubbed together and filtered to specifically solve the defined logic.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml
<Document>
        <name>Employee_Address.kml</name>
          <Placemark>
            <name>ComTech15601_Charles_Mannigan</name>
                        <open>1</open>
                        <LookAt>

            <longitude>73.75097823091289</longitude>

            <latitude>18.60311484559574</latitude>

<heading>-1.033936525450842</heading>

            <range>315.5434276253114</range>

            <gx:altitudeMode>relativeToSeaFloor</gx:altitude
Mode>
                        <Salary>38,000</ Salary >
                    </LookAt>
          </Placemark>
</Document>
</kml>
```

FME can read Excel data from and filter it to make the most of dataset. Sheets to Read Panel is used to read a worksheet and define the row containing the Field Names we wish to use, and the Cell Ranges from which to read the data. The first few rows of the sheet (or named range) that have been selected will appear in the Preview panel. Any changes made to the Sheets to Read panel will be reflected here. Attribute names (Columns) are shown in bold text. All changes to Attribute (Column) formatting can be made by interacting with the Attributes panel. FME will automatically assign types to the Attributes as the data is being read; however, each of the Attributes can also be set manually. FME attempts to create point geometry for features (rows) that appear to reference a coordinate system. Coordinates set manually through the use of the "x_coordinate" and "y_coordinate" data types. FME will read Excel fields that are formatted as a date type and convert them to an FME date string. FME will preserve hyperlinks being read from the source Excel file and preserve the mail information in its original formats.

Feature Paths and flattening is used to convert KML element into an FME feature. Feature paths is used to query KML by defining the node in the KML structure from which extract features. Flattening converts the nested structures within the selected KML element into fields in the form of parent. child. Parent ids can be recorded so you can build associations. This approach replaces the need to use scripts (xfmaps) or text processing to read XML. There is also a tree control that helps define feature paths automatically.

The attribute section confirms that FME has picked the correct column names and properly identified the longitude as type "x_coordinate" and latitude as type "y_coordinate"; When the data is read, the points will automatically be created. ESRI Shape for the format is selected and name is place into the field until it appears within the dropdown list. A shape file of addresses with defined latitude and longitude and selected attribute information is comes as an output. Output file can be viewed in any spatial data visualization framework. An instance of the shape file with defined information is presented in the figure 3.

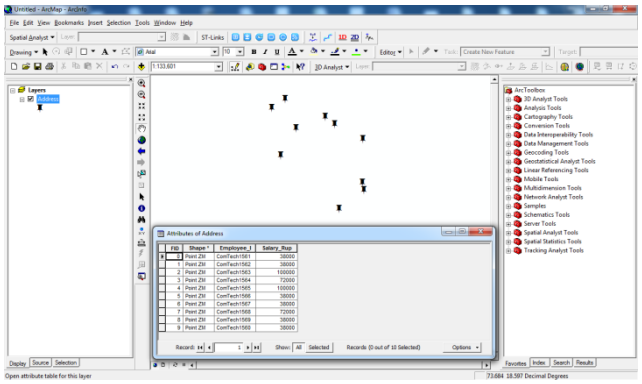

Fig. 3 Shape file showing employee addresses and corresponding attribute information

7.4 Discussion

Implementation strategy helps to filter employee personal information (attribute) and home locations (spatial) information and load essential attributes in the form of shp file. As shown in the figure 3 shp file not only contain location information but also have integrated attribute information. This information is used by company to decide the travelling allowance as per the employee basic salary and his/her home distance from company office. Distance of employee home is computed using network analyses tool. Network analysis tool uses Dijkstra's shortest distance algorithm to select path with minimum weight as distance (See Table 2). Now employee allowance function uses shortest distance and salary as the primary parameter to compute the travelling allowances.

Table 2. Distance from each employee house to company house

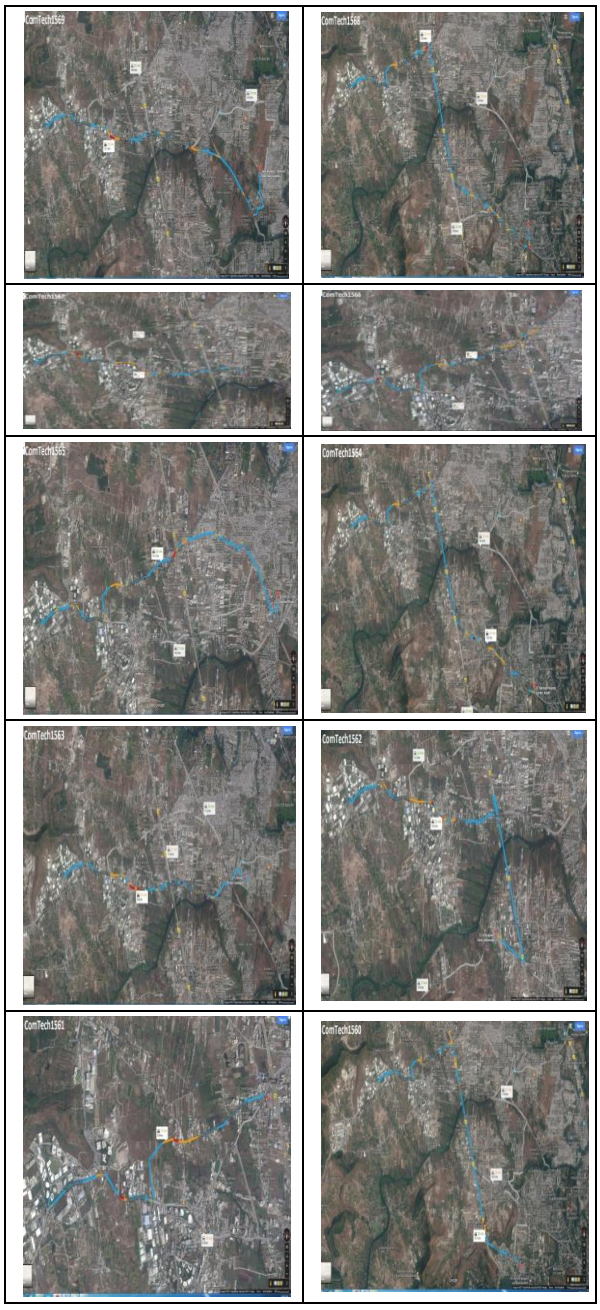

Table 3. Salary and distance (Km) corresponds to each employee

Employee Id	Distance (m)	Salary
ComTech1561	05500	38000
ComTech1562	10300	38000
ComTech1563	09800	100000
ComTech1564	14400	72000
ComTech1565	10000	100000
ComTech1566	06800	38000
ComTech1567	07100	38000
ComTech1568	14000	72000
ComTech1569	13100	38000
ComTech1560	15100	38000

On the basis of computed shortest distance and salary a travelling allowance function is written to quantify the travelling allowances for each employee. This function has predefined weightage for salary and shortest distance. As per the weightage salary and shortest distance were added to obtain travelling allowance. Resultant table will contain Employee id, Shortest Distance, Salary (See Table 3) and travelling allowance corresponds to each employee.

Function trav_alw(Salary, shrt_Dist)

alw_emp = Salary * W1 + shrt_Dist * W2

return alw_emp

Following the above section of the code the company ComTech calculates employee travelling allowances with the 2% weightage of the salary and 10% weightage to the shortest route distance. Thus for the employee with employee id ComTech1564 having monthly travelling allowance of 2880/-. Similarly the formula is used to calculate monthly travelling allowances for all the employees and presented in table 4.

Table 4. Monthly traveling allowance for each employee

Employee Id	Monthly Traveling Allowance
ComTech1561	1310
ComTech1562	1790
ComTech1563	2980
ComTech1564	2880
ComTech1565	3000
ComTech1566	1440
ComTech1567	1470
ComTech1568	2840
ComTech1569	2070
ComTech1560	2270

8. CONCLUSION

As organizations evolve, they acquire or inherit various systems to help the company manage and run their businesses: employee management point-of-sale, inventory management, production control, and general ledger systems—the list can go on and on. Even worse, not only are the systems separated and acquired at different times, but frequently they are logically and physically incompatible. The ETL process needs to effectively integrate systems to achieve the best decisions. It is the best technique to solve the problems having dynamic attribute like employee salary and house distance from company office. Growth of employee designation often changes his/her salary and any change in employee residence also changes his/her shortest travelling distance from employee office. These continuously changing values are required to be fetched at the instance when change occurs.

9. REFERENCES

[1] Jukic, N., 2006. Modeling strategies and alternatives for data warehousing projects. Communications of the ACM, 49(4), pp.83-88.

[2] Kimball, R. and Ross, M., 2011. The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

[3] Cuzzocrea, A., Song, I.Y. and Davis, K.C., 2011, October. Analytics over large-scale multidimensional data: the big data revolution!. In Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP (pp. 101-104). ACM.

[4] Eckerson, W. and White, C., 2003. Evaluating ETL and data integration platforms. Seattle: The DW Institute.

[5] Golfarelli, M., Rizzi, S. and Cella, I., 2004, November. Beyond data warehousing: what's next in business intelligence?. In Proceedings of the 7th ACM international workshop on Data warehousing and OLAP (pp. 1-6). ACM.

[6] Karakasidis, A., Vassiliadis, P. and Pitoura, E., 2005, June. ETL queues for active data warehousing. In Proceedings of the 2nd international workshop on Information quality in information systems (pp. 28-39). ACM.

[7] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P., 2003. Fundamentals of Data Warehouses, second ed. Springer-Verlag.

[8] Suresh, S., Gautam, J.P., Pancha, G., DeRose, F.J. and Sankaran, M., Informatica Corporation, 2001. Method and architecture for automated optimization of ETL throughput in data warehousing applications. U.S. Patent 6,208,990.

[9] Berson, A., Smith, S.J., 1997. Data Warehousing, Data Mining, and OLAP. McGraw-Hill.

[10] Moss, L.T., 2005. Moving Your ETL Process into Primetime. (visited June 2005).

[11] Feature Manipulation Engine (FME): https://www.safe.com

[12] Microsoft Office Home: https://www.office.com

[13] Google Earth: http://earth.google.com

[14] KML OGC:www.opengeospatial.org/standards/ kml

10. APPENDIX

Table 5. Employees attribute information for selected employees

Employee Id	First Name	Last Name	Designation	DOB	Date of Joining	Salary (Rupees)	Email
ComT1561	Charles	Mannigan	Coordinator	29/04/1980	25/11/2015	38,000/-	ChaCoo@CoT.in
ComT1562	Janine	Keys	Coordinator	20/08/1983	25/11/2015	38000/-	JanCoo@CoT.in
ComT1563	Brock	Henderson	Senior Manager	14/07/1969	15/09/2012	10,00,00/-	BroSen@CoT.in
ComT1564	Horace	Shackely	Manager	02/09/1971	02/11/2013	72000/-	HorMan@CoT.in
ComT1565	Ryan	Baxter	Senior Manager	26/03/1974	13/09/2012	10,00,00/-	RyaSen@CoT.in
ComT1566	Sarah	Schrek	Coordinator	18/12/1981	19/05/2016	38000/-	SarCoo@CTe.in
ComT1567	Tanner	Kendrick	Coordinator	30/10/1979	19/05/2016	38000/-	TanCoo@CoT.in
ComT1568	Rebecca	Hart	Manager	03/07/1972	14/02/2013	72000/-	RebMan@CoT.in
ComT1569	Normal	Coleman	Coordinator	15/05/1980	30/07/2016	38000/-	NorCoo@CoT.in
ComT1560	Sarfraz	Ali	Coordinator	05/09/1982	25/11/2015	38000/-	SarCoo@CoT.in